Statistica Sinica **28** (2018), 255-276
doi:https://doi.org/10.5705/ss.202015.0202

INFERENCE OF HIGH-DIMENSIONAL LINEAR MODELS WITH TIME-VARYING COEFFICIENTS

Xiaohui Chen and Yifeng He

University of Illinois at Urbana-Champaign

Abstract: We propose a pointwise inference algorithm for high-dimensional linear models with time-varying coefficients. The method is based on a novel combination of the nonparametric kernel smoothing technique and a Lasso bias-corrected ridge regression estimator. Due to the non-stationarity feature of the model, dynamic bias-variance decomposition of the estimator is obtained. With a bias-correction procedure, the local null distribution of the estimator of the time-varying coefficient vector is characterized for iid Gaussian and heavy-tailed errors. The limiting null distribution is also established for Gaussian process errors, and we show that the asymptotic properties differ between short-range and long-range dependent errors. Here, p-values are adjusted by a Bonferroni-type correction procedure to control the familywise error rate (FWER) in the asymptotic sense at each time point. The finite sample size performance of the proposed inference algorithm is illustrated with synthetic data and an application to learn brain connectivity by using the resting-state fMRI data for Parkinson's disease.

Key words and phrases: Asymptotic theory, high-dimensional linear models, statistical inference, time-varying coefficients, time series analysis.

1. Introduction

We consider the time-varying coefficient models (TVCM)

$$y(t) = \mathbf{x}(t)^\top \boldsymbol{\beta}(t) + e(t), \tag{1.1}$$

where $t \in [0,1]$ is the time index, $y(\cdot)$ the response process, $\mathbf{x}(\cdot)$ the $p \times 1$ deterministic predictor process, $\boldsymbol{\beta}(\cdot)$ the $p \times 1$ time varying coefficient vector, and $e(\cdot)$ the mean zero stationary error process. The response and predictors are observed at $t_i = i/n, i = 1, ..., n$, i.e. $y_i = y(t_i), \mathbf{x}_i = \mathbf{x}(t_i)$, and $e_i = e(t_i)$ with a known covariance matrix $\Sigma_e = \mathrm{Cov}(\mathbf{e})$ where $\mathbf{e} = (e_1, \cdots, e_n)^\top$. TVCM is useful for capturing the dynamic associations in the regression models and longitudinal data analysis Hoover et al. (1998), and it has broad applications in biomedical engineering, environmental science, and econometrics. In this paper, we focus on the *pointwise* inference for the time-varying coefficient vector $\boldsymbol{\beta}(t)$ in the high-dimensional double asymptotic framework $\min(p, n) \to \infty$.

Nonparametric estimation and inference of the TVCM in fixed dimension has been extensively studied, see e.g. Robinson (1989); Hoover et al. (1998); Cleveland, Grosse and Shyu (1991); Fan and Zhang (1999); Zhang, Lee and Song (2002); Orbe, Ferreira and Rodriguez-Poo (2005); Cai (2007); Zhang and Wu (2012); Zhou and Wu (2010). In high dimensions, variable selection and estimation of varying-coefficient models using basis expansions have been studied in Wei, Huang and Li (2011); Xue and Qu (2012); Song, Yi and Zou (2014). Our primary objective is not to estimate $\boldsymbol{\beta}(t)$, but rather to perform statistical inference on the coefficients. In particular, for any $t \in (0,1)$, we wish to test the local hypothesis, for $j = 1, \cdots, p$,

$$H_{0,j,t} : \beta_j(t) = 0 \quad \text{VS} \quad H_{1,j,t} : \beta_j(t) \neq 0. \tag{1.2}$$

By assigning p-values at each time point, we construct a sequence of estimators of the coefficient vectors that allows us to assess the uncertainty of the dynamic patterns in such as brain connectivity networks. Confidence intervals and hypothesis testing problems of lower-dimensional functionals of the high-dimensional constant coefficient vector $\boldsymbol{\beta}(t) \equiv \boldsymbol{\beta}, \forall t \in [0,1]$, have been studied in Bühlmann (2013); Zhang and Zhang (2013); Javanmard and Montanari (2014). To the best of our knowledge, little has been done for inference of the high-dimensional TVCM and our goal is to fill the inference gap between the classical TVCM and the high-dimensional linear model.

While the existing literature on high-dimensional linear models is based on iid errors, (Bühlmann (2013); Zhang and Zhang (2013); Javanmard and Montanari (2014)), we provide an asymptotic theory for answering the question that to which extent the statistical validity of inferences based on iid errors can hold for dependent errors. Allowing temporal dependence is of the practical interest as many datasets such as fMRI data are spatio-temporal and the errors are naturally correlated in the time domain. Theoretical analysis has revealed that the temporal dependence has delicate impact on the asymptotic rates for estimating the covariance structures, Chen, Xu and Wu (2013, 2016). Therefore, it is useful to build an inference procedure that is also robust in the time series context. The error process e_i is modelled as a stationary linear process

$$e_i = \sum_{m=0}^{\infty} a_m \xi_{i-m}, \tag{1.3}$$

where $a_0 = 1$ and ξ_i are iid mean-zero random variables (a.k.a. innovations) with variance σ^2. When the ξ_i are normal, the linear processes of form (1.3) are Gaussian processes that cover the autoregressive and moving-average (ARMA) models with iid Gaussian innovations as special cases. For the linear process, we deal with both weak and strong temporal dependences. In particular, if $a_m = O((m+1)^{-\varrho})$ and $\varrho > 1/2$, then e_i is well-defined and has (i) short-range dependence (SRD) if $\varrho > 1$, (ii) long-range dependence (LRD) if $1/2 < \varrho < 1$. For the SRD processes, it is clear that $\sum_{m=0}^{\infty} |a_m| < \infty$ and therefore the long-run variance is finite.

The paper is organized as follows. In Section 2, we describe our method in details. Asymptotic theory is presented in Section 3. Section 4 presents some simulation results and Section 5 demonstrates an application to an fMRI dataset. The paper concludes in Section 6 with a discussion of some future work. Proofs and some implementation issues are available in the Appendix.

2. Method

2.1. Notations and preliminary

Let K be a non-negative symmetric function with bounded support in $[-1, 1]$, $\int_{-1}^{1} K(x)dx = 1$, and let b_n be a bandwidth parameter satisfying $b_n = o(1)$ and $n^{-1} = o(b_n)$. For each time point $t \in \varpi = [b_n, 1-b_n]$, the Nadaraya-Waston smoothing weight is defined as

$$w(i,t) = \begin{cases} \dfrac{K_{b_n}(t_i - t)}{\sum_{m=1}^{n} K_{b_n}(t_m - t)} & \text{if } |t_i - t| \le b_n, \\ 0 & \text{otherwise,} \end{cases} \tag{2.1}$$

where $K_b(\cdot) = K(\cdot/b)$. Let $N_t = \{i : |t_i - t| \le b_n\}$ be the b_n-neighborhood of time t, $|N_t|$ be the cardinality of the discrete set N_t, $W_t = \mathrm{diag}(w(i,t)_{i \in N_t})$ be the $|N_t| \times |N_t|$ diagonal matrix with $w(i,t), i \in N_t$ on the diagonal, and let $\mathcal{R}_t = \mathrm{span}(\mathbf{x}_i : i \in N_t)$ be the subspace in $\mathbb{R}^p$ spanned by $\mathbf{x}_i$, the rows of design matrix X in the N_t neighborhood. Let $\mathcal{X}_t = (w(i,t)^{1/2}\mathbf{x}_i)_{i \in N_t}^{\top}$, $\mathcal{Y}_t = (w(i,t)^{1/2} y_i)_{i \in N_t}^{\top}$, and $\mathcal{E}_t = (w(i,t)^{1/2} e_i)_{i \in N_t}^{\top}$. Denote I_p as the $p \times p$ identity matrix. We write the singular value decomposition (SVD) of $\mathcal{X}_t$ as

$$\mathcal{X}_t = PDQ^{\top}, \tag{2.2}$$

where P and Q are $|N_t| \times r$, and $p \times r$ matrices such that $P^{\top}P = Q^{\top}Q = I_r$, and $D = \mathrm{diag}(d_1, \cdots, d_r)$ is a diagonal matrix containing the r nonzero singular values of $\mathcal{X}_t$. Now let $P_{\mathcal{R}_t}$ be the projection matrix onto $\mathcal{R}_t$,

$$P_{\mathcal{R}_t} = \mathcal{X}_t^{\top}(\mathcal{X}_t\mathcal{X}_t^{\top})^{-}\mathcal{X}_t = QQ^{\top}, \tag{2.3}$$

where $(\mathcal{X}_t\mathcal{X}_t^{\top})^{-} = PD^{-2}P^{\top}$ is the pseudo-inverse matrix of $\mathcal{X}_t\mathcal{X}_t^{\top}$. Let $\boldsymbol{\theta}(t) = P_{\mathcal{R}_t}\boldsymbol{\beta}(t)$ be the projection of $\boldsymbol{\beta}(t)$ onto $\mathcal{R}_t$ such that $B(t) = \boldsymbol{\theta}(t) - \boldsymbol{\beta}(t)$ is the projection bias. Let

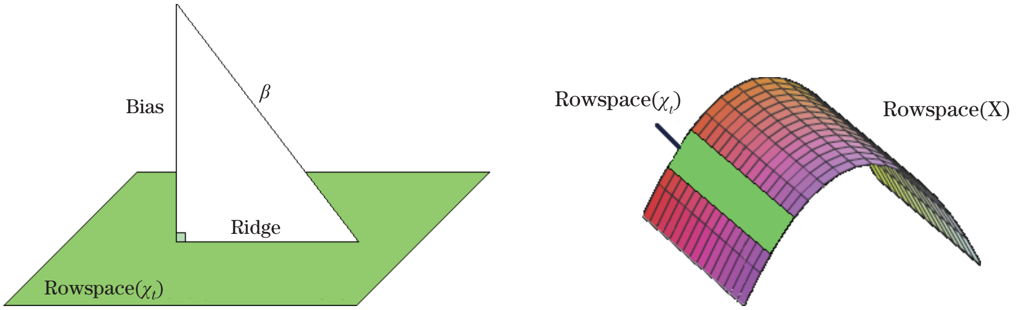
$$\Omega(\lambda) = (\mathcal{X}_t^\top \mathcal{X}_t + \lambda I_p)^{-1} \mathcal{X}_t^\top W_t^{1/2} \Sigma_{e,t} W_t^{1/2} \mathcal{X}_t (\mathcal{X}_t^\top \mathcal{X}_t + \lambda I_p)^{-1} \tag{2.4}$$

be the covariance matrix of the time-varying ridge estimator defined in (2.6), where $\Sigma_{e,t} = \mathrm{Cov}((e_i)_{i \in N_t})$ and $\lambda > 0$ is the shrinkage parameter of the ridge estimator. Let $\Omega_{\min}(\lambda) = \min_{j \le p} \Omega_{jj}(\lambda)$ be the smallest diagonal entry of $\Omega(\lambda)$. For a generic vector $\mathbf{b} \in \mathbb{R}^p$, we write $|\mathbf{b}|_q = (\sum_{j=1}^p |b_j|^q)^{1/q}$ if $q > 0$, and $|\mathbf{b}|_0 = \sum_{j=1}^p \mathbf{1}(b_j \neq 0)$ if $q = 0$. Let $\underline{w}_t = \inf_{i \in N_t} w(i,t)$ and $\overline{w}_t = \sup_{i \in N_t} w(i,t)$. For an $n \times n$ square symmetric matrix M and an $n \times m$ rectangle matrix R, we use $\rho_i(M)$ and $\sigma_i(R)$ to denote the i-th largest eigenvalues of M and singular values of R, respectively. If $k = \mathrm{rank}(R)$, then $\sigma_1(R) \ge \sigma_2(R) \ge \cdots \ge \sigma_k(R) > 0 = \sigma_{k+1}(R) = \cdots = \sigma_{\max(m,n)}(R)$, zeros being padded to the last $\max(m,n) - k$ singular values. We take $\rho_{\max}(M)$, $\rho_{\min}(M)$ and $\rho_{\min \neq 0}(M)$ as the maximum, minimum and nonzero minimum eigenvalues of M, respectively, and $|M|_\infty = \max_{1 \le j,k \le p} |M_{jk}|$. Let

$$\rho_{\max}(M, s) = \max_{|\mathbf{b}|_0 \le s, \mathbf{b} \neq \mathbf{0}} \frac{\mathbf{b}^\top M \mathbf{b}}{\mathbf{b}^\top \mathbf{b}}.$$

If M is nonnegative definite, then $\rho_{\max}(M, s)$ is the restricted maximum eigenvalues of M at most s columns and rows.

The p-dimensional coefficient vector $\boldsymbol{\beta}(t)$ is decomposed into two parts via projecting onto the $|N_t|$-dimensional linear subspace spanned by the rows of $\mathcal{X}_t$ and its orthogonal complement; see Figure 1(a). A key advantage of this decomposition is that the projected part can be conveniently estimated in closed-form, for example, by the ridge estimator since it lies in the row space of $\mathcal{X}_t$ and thus is amenable for the subsequent inferential analysis. In the high-dimensional situation, this projection introduces a non-negligible shrinkage bias in estimating $\boldsymbol{\beta}(t)$ and therefore we may lose information because $p \gg |N_t|$. On the other hand, the shrinkage bias can be corrected by a consistent estimator of $\boldsymbol{\beta}(t)$. As a particular example, we use the Lasso estimator, though any sparsity-promoting estimator attaining the same convergence rate as the Lasso should work. Because of the time-varying nature of the nonzero functional $\boldsymbol{\beta}(t)$, the smoothness on the row space of $\mathcal{X}_t$ along the time index t is necessary to apply nonparametric smoothing technique; see Fig. 1(b). As a special case, when the nonzero components $\boldsymbol{\beta}(t) \equiv \boldsymbol{\beta}$ are constant functions and the error process is iid Gaussian, our algorithm is the same as that of Bühlmann (2013). Here, we emphasize that (i) coefficient vectors are time-varying (i.e. non-constant), (ii) errors are allowed to have heavy-tails by assuming milder polynomial moment conditions and to have



(a) Bias correction by projection to the row space of $\mathcal{X}_t$.

(b) Smoothly time-varying row space of $\mathcal{X}_t$.

Figure 1. Intuition of the proposed algorithm in Section 2.2.

temporal dependence, including both SRD and LRD processes. There are other inferential methods for high-dimensional linear models such as Zhang and Zhang (2013); Javanmard and Montanari (2014). We do not explore specific choices here since our contribution is a general framework of combining nonparametric smoothing and bias-correction methods to make inference for high-dimensional TVCM. However, we expect that a non-stationary generalization would be feasible for those methods as well. Some simulation comparisons are provided for time-varying versions of the bias-correction methods in Section 4.

2.2. Inference algorithm

First, we estimate the projection bias $B(t)$ by $\tilde{B}(t) = (P_{\mathcal{R}_t} - I_p)\tilde{\boldsymbol{\beta}}(t)$, where $\tilde{\boldsymbol{\beta}}(t)$ is the time-varying Lasso (tv-Lasso) estimator

$$\begin{aligned} \tilde{\boldsymbol{\beta}}(t) &= \arg\min_{\mathbf{b}\in\mathbb{R}^p} \sum_{i\in N_t} w(i,t)(y_i - \mathbf{x}_i^\top \mathbf{b})^2 + \lambda_1 |\mathbf{b}|_1 \qquad (2.5) \\ &= \arg\min_{\mathbf{b}\in\mathbb{R}^p} |\mathcal{Y}_t - \mathcal{X}_t \mathbf{b}|_2^2 + \lambda_1 |\mathbf{b}|_1. \end{aligned}$$

Next, we estimate $\boldsymbol{\theta}(t) = P_{\mathcal{R}_t}\boldsymbol{\beta}(t)$ using the time-varying ridge (tv-ridge) estimator

$$\begin{aligned} \tilde{\boldsymbol{\theta}}(t) &= \arg\min_{\mathbf{b}\in\mathbb{R}^p} \sum_{i\in N_t} w(i,t)(y_i - \mathbf{x}_i^\top \mathbf{b})^2 + \lambda_2 |\mathbf{b}|_2^2 \\ &= (\mathcal{X}_t^\top \mathcal{X}_t + \lambda_2 I_p)^{-1} \mathcal{X}_t^\top \mathcal{Y}_t. \qquad (2.6) \end{aligned}$$

We defer the discussion of tuning parameters choice λ_1 and λ_2 to Section 3. Our tv-Lasso bias-corrected tv-ridge regression estimator for $\boldsymbol{\beta}(t)$ is

$$\hat{\boldsymbol{\beta}}(t) = \tilde{\boldsymbol{\theta}}(t) - \tilde{B}(t). \qquad (2.7)$$

Based on $\hat{\boldsymbol{\beta}}(t) = (\hat{\beta}_1(t), \cdots, \hat{\beta}_p(t))^\top$, we calculate the raw two-sided p-values for individual coefficients

$$\tilde{P}_j = 2\left[1-\Phi\left(\frac{|\hat{\beta}_j(t)| - \lambda_1^{1-\xi}\max_{k\neq j}|(P_{\mathcal{R}_t})_{jk}|}{\Omega_{jj}^{1/2}(\lambda_2)}\right)\right], \quad j=1,\cdots,p, \tag{2.8}$$

where $\xi \in [0,1)$ is user pre-specified number that depends on the number of nonzero $\boldsymbol{\beta}(t)$. In particular, if $|\text{supp}(\boldsymbol{\beta}(t))|$ is bounded, then we can choose $\xi = 0$. Generally, following Bühlmann (2013), we use $\xi = 0.05$ in our numeric examples to allow the number of nonzero components in $\boldsymbol{\beta}(t)$ to diverge at proper rates. Let $\mathbf{v}(t) = (V_1(t),\cdots,V_p(t))^\top \sim N(\mathbf{0},\Omega(\lambda_2))$ and define the distribution function

$$F(z) = \mathbb{P}\left(\min_{j\le p} 2\left[1-\Phi\left(\Omega_{jj}^{-1/2}(\lambda_2)|V_j(t)|\right)\right] \le z\right). \tag{2.9}$$

We adjust the $\tilde{P}_j$ for multiplicity by $P_j = F(\tilde{P}_j+\zeta)$, where ζ is another pre-defined small number (Bühlmann (2013)) that accommodates asymptotic approximation errors. Our decision rule is defined as: reject $H_{0,j,t}$ if $P_j \le \alpha$ for $\alpha \in (0,1)$. For iid errors, since $\Sigma_e = \sigma^2 I_n$ and

$$\Omega(\lambda_2) = \sigma^2(\mathcal{X}_t^\top\mathcal{X}_t + \lambda_2 I_p)^{-1}\mathcal{X}_t^\top W_t \mathcal{X}_t(\mathcal{X}_t^\top\mathcal{X}_t + \lambda_2 I_p)^{-1},$$

we see that $F(\cdot)$ is independent of σ. Therefore, $F(\cdot)$ can be easily estimated by repeatedly sampling from the multivariate Gaussian distribution $N(\mathbf{0},\Omega(\lambda_2))$. A similar observation has been made in Bühlmann (2013).

3. Asymptotic Results

In this section, we present the asymptotic theory of the inference algorithm in Section 2.2. First, we state the main assumptions for iid Gaussian errors.

1. **Error.** The errors $e_i \sim N(0,\sigma^2)$ are independent and identically distributed (iid).

2. **Sparsity.** $\boldsymbol{\beta}(\cdot)$ is uniformly s-sparse, i.e. $\sup_{t\in[0,1]}|S_t^*| \le s$, where $S_t^* = \{j : \beta_j(t) \neq 0\}$ is the support set.

3. **Smoothness.**

   (a) $\boldsymbol{\beta}(\cdot)$ is twice differentiable with bounded and continuous first and second derivatives in the coordinatewise sense, i.e. $\beta_j(\cdot) \in \mathcal{C}^2([0,1],C_0)$ for each $j = 1,\cdots,p$ and C_0 is an upper bound for the partial derivatives.

   (b) The b_n-neighborhood covariance matrix $\hat{\Sigma}_t^\diamond = |N_t|^{-1}\sum_{i\in N_t}\mathbf{x}_i\mathbf{x}_i^\top := \mathcal{X}_t^{\diamond\top}\mathcal{X}_t^\diamond$ satisfies

$$\rho_{\max}(\hat{\Sigma}_t^\diamond, s) \le \varepsilon_0^{-2} < \infty. \tag{3.1}$$

4. **Non-degeneracy.**
$$\liminf_{\lambda \downarrow 0} \Omega_{\min}(\lambda) > 0. \tag{3.2}$$

5. **Identifiability.**
   (a) The minimum nonzero eigenvalue condition
   $$\rho_{\min \neq 0}(\hat{\Sigma}_t^{\diamond}) \geq \varepsilon_0^2 > 0. \tag{3.3}$$
   (b) The *restricted eigenvalue condition*
   $$\phi_0 = \inf \left\{ \phi > 0 : \min_{|S|=s} \inf_{|\mathbf{b}_{S^c}|_1 \leq 3|\mathbf{b}_S|_1} \frac{\mathbf{b}^\top \hat{\Sigma}_t \mathbf{b}}{|\mathbf{b}_S|_1^2} \geq \frac{\phi^2}{s} \right.$$
   $$\left. \text{holds for all } t \in [0,1] \right\} > 0, \tag{3.4}$$
   where $\hat{\Sigma}_t = \mathcal{X}_t^\top \mathcal{X}_t$ is the kernel smoothed covariance matrix of the predictors.

6. **Kernel.** The kernel function $K(\cdot)$ is nonnegative, symmetric around 0 with bounded support in $[-1, 1]$.

Here, we comment the assumptions and their implications. Assumption 1 and 6 are standard. The Gaussian distribution is non-essential and can be relaxed to sub-Gaussian and heavier tailed distributions (Theorem 4). Assumption 2 is a sparsity condition for the nonzero functional components and allows that $s \to \infty$ slower than $\min(p, n)$. It is a key condition for maintaining the low-dimensional structure when the dimension p grows with the sample size n. By the argument of Theorem 5 in Zhou, Lafferty and Wasserman (2010), it implies that the number of the first and second non-vanishing derivatives of $\boldsymbol{\beta}(t)$ is bounded by s almost surely on $[0, 1]$. Assumption 3 ensures the smoothness of the time-varying coefficient vectors and the design matrix so that nonparametric smoothing techniques are applicable. Examples of Assumption 3(a) include the quadratic functions $\boldsymbol{\beta}(t) = \boldsymbol{\beta} + \boldsymbol{\alpha} t + \boldsymbol{\xi} t^2/2$ and the periodic functions $\boldsymbol{\beta}(t) = \boldsymbol{\beta} + \boldsymbol{\alpha} \sin(t) + \boldsymbol{\xi} \cos(t)$ with $|\boldsymbol{\alpha}|_\infty + |\boldsymbol{\xi}|_\infty \leq C_0$. Assamption 3(b) can be viewed as Lipschitz continuity on the local design matrix that is smoothly evolving, Zhou and Wu (2010). It is weaker than the condition that $\rho_{\max}(\hat{\Sigma}_t^{\diamond}) \leq \varepsilon_0^{-2}$ because the latter may grow to infinity much faster than the restricted form (3.1). Assumption 4 is required for a non-degenerated stochastic component of the proposed estimator which is used for the inference purpose. Assumption 5(a) and 5(b) together impose the identifiability conditions for recovering the coefficient vectors. The analogous condition of the time-invariant version has been extensively used in literature to derive theoretical properties of the Lasso model; see e.g. Bickel, Ritov and Tsybakov (2009); van de Geer and Bühlmann (2009).

For the tv-lasso bias-corrected tv-ridge estimator (2.7), we have the following representation theorem.

Theorem 1 (Representation). *Fix $t \in \varpi$ and let*

$$L_{t,\ell} = \max_{j \le p} \left[\sum_{i \in N_t} w(i,t)^{\ell} X_{ij}^2 \right]^{1/2}, \ \ell = 1, 2, \cdots, \quad \lambda_0 = 4\sigma L_{t,2} \sqrt{\log p}, \tag{3.5}$$

and $\lambda_1 \ge 2(\lambda_0 + 2C_0 L_{t,1} b_n (s|N_t|\overline{w}_t)^{1/2} \varepsilon_0^{-1})$. If $\lambda_2 = o(1)$, Assumptions 1-6 hold, and $C \le |N_t|\underline{w}_t \le |N_t|\overline{w}_t \le C^{-1}$ for some $C \in (0,1)$, then $\hat{\boldsymbol{\beta}}(t)$ admits the decomposition

$$\hat{\boldsymbol{\beta}}(t) = \boldsymbol{\beta}(t) + \mathbf{z}(t) + \boldsymbol{\gamma}(t), \tag{3.6}$$

$$\mathbf{z}(t) \sim N(\mathbf{0}, \Omega(\lambda_2)), \tag{3.7}$$

$$|\gamma_j(t)| \le \frac{\lambda_2 |\boldsymbol{\theta}(t)|_2 + 2C_0 s^{1/2} b_n}{C\varepsilon_0^2} + \frac{4\lambda_1 s}{\phi_0^2} |P_{\mathcal{R}_t} - I_p|_\infty, \quad j = 1, \cdots, p, \tag{3.8}$$

with probability tending to one. If $\beta_j(t) = 0$, then we have

$$\Omega_{jj}^{-1/2}(\lambda_2)(\hat{\beta}_j(t) - \gamma_j(t)) \sim N(0,1), \tag{3.9}$$

where

$$|\gamma_j(t)| \le \frac{\lambda_2 |\boldsymbol{\theta}(t)|_2 + 2C_0 s^{1/2} b_n}{C\varepsilon_0^2} + \frac{4\lambda_1 s}{\phi_0^2} \max_{k \ne j} |(P_{\mathcal{R}_t})_{jk}| \tag{3.10}$$

with probability tending to one.

Remark 1. The decomposition (3.6) can be viewed as a *local version* of the one proposed in Bühlmann (2013) (Proposition 2). However, due to the time-varying nature of the nonzero coefficient vectors, both the stochastic component $\mathbf{z}(t)$ in (3.7) and the bias component $\boldsymbol{\gamma}(t)$ in (3.8) differ from Bühlmann (2013). First, our bound (3.8) for bias has three terms arising from: ridge shrinkage, *non-stationarity* and Lasso correction, and each has localized features depending on the bandwidth b_n of the sliding window and the smoothness parameter C_0. Second, the stochastic part (3.7) also has time-dependent features in the covariance matrix (i.e. $\Omega(\lambda_2)$ implicitly depends on t though $\mathcal{X}_t$) and the scale of normal random vector is different from Bühlmann (2013). Delicate balance among them allows us to perform valid statistical inference such as hypothesis testing and confidence interval construction for the coefficients and, more broadly, their lower-dimensional linear functionals.

Example 1. Consider the uniform kernel $K(x) = 0.5\mathbb{I}(|x| \le 1)$ as an important special case, the kernel used for our numeric experiments in Section 4. In this case, $\underline{w}_t = (2nb_n)^{-1}$ and $|N_t|\underline{w}_t = |N_t|\overline{w}_t = 1$. It is easily verified that under the local null hypothesis $H_{0,j,t}$, (3.10) can be simplified to

$$\gamma_j(t) = O\left(\lambda_2|\boldsymbol{\theta}(t)|_2 + s^{1/2}b_n + \lambda_1 s \max_{k\neq j} |(P_{\mathcal{R}_t})_{jk}|\right).$$

From this, it is clear that the three terms correspond to bias of ridge-shrinkage, non-stationarity and Lasso-correction. The first and last components have dynamic features and the non-stationary bias is controlled by the bandwidth and sparsity parameters. The condition $C \le |N_t|\underline{w}_t \le |N_t|\overline{w}_t \le C^{-1}$ in Theorem 1 rules out the case that the kernel does not use the boundary rows in the localized window and therefore avoids any jump in the time-dependent row subspaces.

Remark 2. In Theorem 1, the penalty level for the tv-Lasso estimator λ_1 can be chosen as $O(\sigma L_{t,2}\sqrt{\log p} + L_{t,1}s^{1/2}b_n)$. The second term in the penalty is due to the non-stationarity of $\boldsymbol{\beta}(t)$ and the factor $s^{1/2}$ arises from the weak coordinatewise smoothness requirement on its derivatives (Assumption 3(a)). In the Lasso case with $\boldsymbol{\beta}(t) \equiv \boldsymbol{\beta}$ and $w(i,t) \equiv n^{-1}$, an ideal order of the penalty level λ_1 is $\sigma n^{-1} \max_{j\le p}(\sum_{i=1}^n X_{ij}^2)^{1/2}(\log p)^{1/2}$ see e.g. Bickel, Ritov and Tsybakov (2009). In the standardized design case $n^{-1}\sum_{i=1}^n X_{ij}^2 = 1$ so that $L_{t,1} = 1$ and $L_{t,2} = n^{-1/2}$, the Lasso penalty is $O(\sigma(n^{-1}\log p)^{1/2})$, while the tv-Lasso has an additional term $s^{1/2}b_n$ that may cause a larger bias. In our case, we estimate the time-varying coefficient vectors by smoothing the data points in the localized window. Thus, it is unnatural to standardize the reweighted local design matrix to have unit ℓ^2 length and the additional bias $O(s^{1/2}b_n)$ is due to non-stationarity. If the X_{ij} are iid Gaussian random variables without standardization and we interpret the linear model as conditional on X, then, under the uniform kernel, we have $L_{t,2}^2 = O_{\mathbb{P}}(\log p/|N_t|)$ and, in the Lasso case, the penalty level is $O(\sigma|N_t|^{-1/2}\log p)$. If $s = O(\log p)$ and $b_n = O((\log p/n)^{1/3})$, then the choice in Theorem 1 has the same order as the Lasso with constant coefficient vector.

Based on Theorem 1, we can prove that the inference algorithm in Section 2.2 asymptotically controls the familywise error rate (FWER). Let $\alpha \in (0,1)$ and $\mathrm{FP}_\alpha(t)$ be the number of false rejections of $H_{0,j,t}$ based on the adjusted p-values. In the asymptotic statement, $p := p(n)$ is a function of n such that $p \to \infty$ as $n \to \infty$.

Theorem 2 (Pointwise inference: multiple testing)**.** *If the conditions of Theorem*

1 *hold and*

$$\lambda_2|\boldsymbol{\theta}(t)|_2 + s^{1/2}b_n = o(\Omega_{\min}(\lambda_2)^{1/2}), \tag{3.11}$$

then we have for each fixed $t \in \varpi$

$$\limsup_{n\to\infty} \mathbb{P}(FP_\alpha(t) > 0) \leq \alpha. \tag{3.12}$$

The proof of Theorem 2 is standard by combining the argument of Theorem 2 in Bühlmann (2013) and Theorem 1. Therefore, we omit the proof. Condition (3.11) essentially requires that the shrinkage and non-stationarity biases of the tv-ridge estimator together are dominated by the variance; see also the representation (3.6), (3.7), (3.8), and (3.9). This is mild condition for two reasons. First, in view that variance of the tv-ridge estimator is lower bounded when λ_2 is small enough; c.f. (3.2), the first term is quite weak in the sense that the tv-ridge estimator acts on a much smaller subspace with dimension $|N_t|$ than the original p-dimensional vector space. Second, for the choice of penalty parameter of λ_1 in Theorem 1, the term $s^{1/2}b_n$ in (3.11) is at most λ_1. Hence, the bias correction (including the projection and non-stationary parts) in the inference algorithm (2.8) has a dominating effect on the second term of (3.11). Consequently, provided λ_2 is small enough, the bias correction step in computing the raw p-value asymptotically approximates the stochastic component in the tv-ridge estimator.

Remark 3. The Bonferroni correction (2.9) for the raw p-values is often conservative and thus it may be sub-optimal in power. In our simulation studies, it seems that detection power is reasonable while the FWER is controlled at 0.05; c.f. Table 1 and 2. To improve the power, one can consider the control of the false discovery proportion (FDP) by the principal factor approximation (PFA) method proposed in Fan, Han and Gu (2012); Fan and Han (2016). By Theorem 1, under the global null hypothesis $H_{0,t} : \beta_1(t) = \cdots = \beta_p(t) = 0$, we have $\hat{\boldsymbol{\beta}}(t) - \boldsymbol{\gamma}(t) \sim N(\mathbf{0}, \Omega(\lambda_2))$ with a known covariance matrix $\Omega(\lambda_2)$. Therefore, our test statistic is jointly normal and the PFA can be applied to control the FDP if $\Omega(\lambda_2)$ can be well approximated by the covariance matrix of a factor model plus a weakly dependent component. Fan, Han and Gu (2012) provided a practical procedure to estimate the FDP.

Next, we relax the iid assumption on the errors to allow temporal dependence.

Theorem 3 (Gaussian process errors)**.** *Suppose that the error process* e_i *is a mean-zero stationary Gaussian process of form* (1.3) *such that* $|a_m| \leq K(m+1)^{-\varrho}$

for some $\varrho \in (1/2, 1) \cup (1, \infty)$ and finite constant $K > 0$. Under Assumptions 2-6 *and the notation of Theorem* 1 *with*

$$\lambda_0 = \begin{cases} 4\sigma L_{t,2}|\mathbf{a}|_1\sqrt{\log p} & \text{if } \varrho > 1, \\ C_{\varrho,K}\sigma L_{t,2}n^{1-\varrho}\sqrt{\log p} & \text{if } 1 > \varrho > 1/2, \end{cases} \tag{3.13}$$

where $\mathbf{a} = (a_0, a_1, \cdots)^\top$, we have the representation of $\hat{\boldsymbol{\beta}}(t)$ in (3.6)–(3.10) *with probability tending to one.*

From Theorem 3, the temporal dependence strength has a dichotomous effect on the choice of λ_0, and therefore on the asymptotic properties of $\hat{\boldsymbol{\beta}}(t)$. For e_i with SRD, we have $|\mathbf{a}|_1 < \infty$ and $\lambda_0 \asymp \sigma L_{t,2}\sqrt{\log p}$. Therefore, the bias-correction part $\boldsymbol{\gamma}(t)$ of estimating $\boldsymbol{\beta}(t)$ has the same rate of convergence as the iid error case. The temporal effect only plays a role in the long-run covariance matrix of the stochastic part $\mathbf{z}(t)$. If e_i has LRD, then the temporal dependence has impact on both $\boldsymbol{\gamma}(t)$ and $\mathbf{z}(t)$. In addition, the choice of the bandwidth parameter b_n is different from the SRD and iid cases. In particular, the optimal bandwidth for $\varrho \in (1/2, 1)$ is $O((\log p/n^{\varrho})^{1/3})$ which is much larger than $O((\log p/n)^{1/3})$ in the iid and SRD cases, assuming s is bounded. The boundary case $\varrho = 1$ can also be characterized; details are omitted.

We also relax the moment condition on the errors that, in the iid error case, are assumed to be zero-mean Gaussian. First, it is easy to relax this assumption to distributions with sub-Gaussian tails (see Definition S0.1 in the Supplementary Material) and Theorem 1 and 2 continue to hold, in view that the large deviation inequality and the Gaussian approximation for a weighted partial sum of the error process only depend on the tail behavior and therefore on moments of e_i. Second and more importantly, the sub-Gaussian assumption may be knocked down to allow iid noise processes with algebraic tails, or equivalently e_i with moments up to a finite order. The consequence of this relaxation is that a larger penalty parameter for the tv-Lasso is needed for errors with polynomial moments. Let Ξ be the square root matrix of $\Omega(\lambda_2)/\sigma^2$ (i.e. $\Omega(\lambda_2) = \sigma^2\Xi\Xi^\top$) and $\boldsymbol{\xi}_j$ be the j-th row of Ξ.

Theorem 4 (Heavy-tailed errors)**.** *Under the conditions of Theorem* 1 *with $\mathbb{E}|e_i|^q < \infty, q > 2$, choose*

$$\lambda_0 = C_q \max\left\{(p\mu_{n,q})^{1/q},\ \sigma L_{t,2}(\log p)^{1/2}\right\}, \quad \text{for large enough } C_q > 0, \tag{3.14}$$

where $\mu_{n,q} = \sum_{i\in N_t}|w(i,t)X_{ij}|^q$. If $|\boldsymbol{\xi}_j|_q = o(|\boldsymbol{\xi}_j|_2)$ for all $j = 1, \cdots, p$, then (3.6) *holds with probability tending to one and Theorem* 2 *holds.*

The assumption $|\boldsymbol{\xi}_j|_q = o(|\boldsymbol{\xi}_j|_2)$ is needed to ensure the asymptotic validity of the Gaussian approximation of the ridge component (3.7) for non-Gaussian data.

4. Simulation Studies

4.1. Simulation setup

In the simulation studies, we generated the $n \times p$ design matrix with iid rows sampled from $N(\mathbf{0}, \Sigma_{\mathbf{X}})$ for $n = 300$ and $p = 300$. We considered two covariance structures on the design matrix: (i) $\Sigma_{\mathbf{X}} = I_p$; (ii) $\Sigma_{\mathbf{X}} = T$, where $T = (t_{jk})_{j,k=1}^{p}$ and $t_{jk} = 0.5^{|j-k|}$. The time-varying coefficient vectors $\boldsymbol{\beta}(t)$ had $s = 3$ non-zero elements and $p - 3$ zeros for all $t \in [0, 1]$. The non-zero elements in $\boldsymbol{\beta}(t)$ were generated by sampling nodes from a uniform distribution $U(-2.5, 2.5)$ at regular time points and smoothly interpolating on the interval $[0, 1]$ using the cubic splines. We simulated the following stationary error processes.

1. The e_i are iid $N(0, 1)$.
2. $e_i = \varphi e_{i-1} + \xi_i$ is an AR(1) process where $\varphi \in \{0.2, 0.5\}$ and ξ_i are iid $N(0, 1)$.
3. The e_i are iid Student's $t(3)/\sqrt{3}$.
4. e_i is a long-memory process $e_i = \sum_{m=0}^{\infty}(m+1)^{-\varrho}\xi_{i-m}$, where $\varrho = 0.75$ and ξ_i are iid Gaussian with mean zero.

We compared the performance of the proposed method with the following.

1. (TV-Lasso) - The time-varying Lasso, the kernel smoothed time-varying LASSO defined in (2.5), where λ_1 is selected by the cross-validation (CV).
2. (FP-Lasso) - The false-positive Lasso, where λ_1 is tuned to match the FWER of the proposed method. This allowed us to compare the power at similar levels of FWER.
3. (TV-LDPE) - An adaptation of the de-biased LASSO inference procedure by Zhang and Zhang (2013) to the kernel smoothed, time-varying setting.
4. (TV-SDL) - An adaptation of the SDL test of Javanmard and Montanari (2014) to the kernel smoothed, time-varying setting.
5. (Non-TV) - The original non-time-varying method of Bühlmann (2013) that ignores the dynamic structures. The penalty parameter λ_1 in the Lasso is set to the scaled Lasso parameter $\sqrt{2\log(p)/n}$.

In all time-varying models, we used the kernel bandwidth $b_n = 0.1$. For the proposed method, we used $\lambda_2 = 1/n$ and $\zeta = 0$. We let $P_{j,t,m}$ be the multiplicity-adjusted p-value for testing the hypothesis $H_{0,j,t} : \beta_j(t) = 0$ for $t \in \varpi = [b_n, 1-b_n]$ in the m-th Monte Carlo simulation for $m = 1, \cdots, M$. For TV-LDPE, TV-SDL, the proposed method and its non-tv version, we adopted the following performance measures.

1. The (averaged) false positive rate (FPR) over the interval ϖ,
$$\frac{1}{n(1-2b_n)(p-s)M} \sum_{t\in\varpi} \sum_{j\in\mathcal{S}^c} \sum_{m=1}^{M} \mathbf{1}(P_{j,t,m} \le \alpha).$$
2. The (averaged) false negative rate (FNR) over the interval ϖ,
$$\frac{1}{n(1-2b_n)sM} \sum_{t\in\varpi} \sum_{j\in\mathcal{S}} \sum_{m=1}^{M} \mathbf{1}(P_{j,t,m} > \alpha).$$
3. The (averaged) FWER over the interval ϖ,
$$\text{FWER} = \frac{1}{n(1-2b_n)M} \sum_{t\in\varpi} \sum_{m=1}^{M} \mathbf{1}(\min_{j\in\mathcal{S}^c} P_{j,t,m} \le \alpha).$$

For the Lasso-based methods (TV-Lasso and FP-Lasso), the probabilities are replaced by the corresponding indicators of whether or not the estimated coefficients are zero.

4.2. Empirical results

For each simulation setup, we report the FPR, RNR, FWER, and the root mean square errors (RMSE) of the estimates. The results are shown in Tables 1 and 2, from which we make several observations. First, TV-Lasso and the method of Bühlmann (2013) do not control the FWER, while the proposed method can control the FWER at the nominal level $\alpha = 0.05$ in all setups. Second, the proposed method has uniformly higher power than FP-Lasso, the TV-Lasso tuned to match the FWER with our method. This is probably explained by the bias of the ℓ_1 regularization in the TV-Lasso. Third, for the design matrix with iid Gaussian entries, TV-LDPE and TV-SDL have comparable performance as our proposed method in terms of the power, while all three methods have the FWER controlled below 0.05. TV-LDPE and TV-SDL are more sensitive to the design matrix than the proposed method; for the Toeplitz design matrix T, TV-LDPE and TV-SDL seem to lose control on the FWER. Moreover, the FWER, FNR, and RMSE are larger as the dependence level of the error process grows and as

Table 1. Simulation results. $n = 300, p = 300, s = 3$.

	$\mathbf{x}_i \overset{\text{iid}}{\sim} N(\mathbf{0}, I_p)$, $e \sim N(\mathbf{0}, I_n)$				$\mathbf{x}_i \overset{\text{iid}}{\sim} N(\mathbf{0}, T)$, $e \sim N(\mathbf{0}, I_n)$			
Method	FPR	FNR	FWER	RMSE	FPR	FNR	FWER	RMSE
TV-Lasso	7.51×10^{-2}	0.0551	1	0.0537	1.55×10^{-1}	0.0684	1	0.0520
FP-Lasso	1.50×10^{-4}	0.2352	0.0344	0.1124	2.81×10^{-5}	0.5170	0.0063	0.1318
Proposed	1.50×10^{-4}	0.1889	0.0346	0.1838	2.81×10^{-5}	0.4072	0.0063	0.1984
TV-LDPE	1.29×10^{-4}	0.1981	0.0254	0.2652	3.77×10^{-4}	0.3615	0.0743	0.2727
TV-SDL	1.53×10^{-4}	0.1848	0.0357	0.1316	1.08×10^{-3}	0.3006	0.2119	0.1409
Non-TV	4.89×10^{-1}	0.5100	0.4600	0.5762	2.58×10^{-1}	0.7100	0.2400	0.6084

	$\mathbf{x}_i \overset{\text{iid}}{\sim} N(\mathbf{0}, I_p)$, $e \sim \text{AR}(1)$ with $\varphi = 0.2$				$\mathbf{x}_i \overset{\text{iid}}{\sim} N(\mathbf{0}, T)$, $e \sim \text{AR}(1)$ with $\varphi = 0.2$			
Method	FPR	FNR	FWER	RMSE	FPR	FNR	FWER	RMSE
TV-Lasso	8.93×10^{-2}	0.0563	1	0.0533	1.51×10^{-1}	0.0629	1	0.0519
FP-Lasso	1.78×10^{-4}	0.2363	0.0384	0.1099	7.21×10^{-5}	0.4709	0.0173	0.1302
Proposed	1.78×10^{-4}	0.1891	0.0376	0.1836	7.21×10^{-5}	0.3434	0.0173	0.1920
TV-LDPE	9.85×10^{-5}	0.1995	0.0215	0.2799	3.70×10^{-4}	0.3620	0.0843	0.2725
TV-SDL	1.97×10^{-4}	0.1883	0.0419	0.1374	1.07×10^{-3}	0.3032	0.2165	0.1404
Non-TV	4.29×10^{-1}	0.5633	0.4100	0.5557	2.58×10^{-1}	0.6933	0.2200	0.6088

	$\mathbf{x}_i \overset{\text{iid}}{\sim} N(\mathbf{0}, I_p)$, $e \sim \text{AR}(1)$ with $\varphi = 0.5$				$\mathbf{x}_i \overset{\text{iid}}{\sim} N(\mathbf{0}, T)$, $e \sim \text{AR}(1)$ with $\varphi = 0.5$			
Method	FPR	FNR	FWER	RMSE	FPR	FNR	FWER	RMSE
TV-Lasso	7.55×10^{-2}	0.0544	1	0.0537	1.51×10^{-1}	0.0611	1	0.0518
FP-Lasso	1.93×10^{-4}	0.2402	0.0431	0.1124	9.81×10^{-5}	0.4805	0.0222	0.1303
Proposed	1.93×10^{-4}	0.1809	0.0422	0.1836	9.81×10^{-5}	0.3347	0.0235	0.1918
TV-LDPE	9.82×10^{-5}	0.2028	0.0229	0.2756	3.71×10^{-4}	0.3618	0.0849	0.2659
TV-SDL	1.94×10^{-4}	0.1898	0.0425	0.1374	1.01×10^{-3}	0.2993	0.2555	0.1439
Non-TV	5.49×10^{-1}	0.4500	0.5200	0.5314	1.70×10^{-1}	0.8300	0.1500	0.6004

the tail of the error distribution becomes thicker. Finally, the proposed method is computationally more economical than the competing methods TV-LDPE and TV-SDL. Table 3 shows the runtimes on an Intel i5-4790K with the Intel MKL linear algebra libraries (software and platform: R 3.2.2 for Windows).

5. Data Example: Learning Brain Connectivity

We illustrate our proposed method in an application to model the functional brain connectivity in a Parkinson's disease study. The problem is to construct brain connectivity networks from the resting-state functional magnetic resonance imaging (fMRI) data, where slowly time-varying graphs have implications in modeling brain connectivity networks. Traditional correlation analysis of the resting state blood-oxygen-level-dependent (BOLD) signals of the brain showed

Table 2. Simulation results (continued). $n = 300, p = 300, s = 3$.

	$\mathbf{x}_i \overset{\text{iid}}{\sim} N(\mathbf{0}, I_p)$, $e_i \overset{\text{iid}}{\sim} t(3)/\sqrt{3}$				$\mathbf{x}_i \overset{\text{iid}}{\sim} N(\mathbf{0}, T)$, $e_i \overset{\text{iid}}{\sim} t(3)/\sqrt{3}$			
Method	FPR	FNR	FWER	RMSE	FPR	FNR	FWER	RMSE
TV-Lasso	9.14×10^{-2}	0.0518	1	0.0539	1.57×10^{-1}	0.0605	1	0.0506
FP-Lasso	1.96×10^{-4}	0.2193	0.0398	0.1120	4.20×10^{-5}	0.4547	0.0124	0.1209
Proposed	1.96×10^{-4}	0.1708	0.0439	0.1834	4.20×10^{-5}	0.3129	0.0125	0.1885
TV-LDPE	1.26×10^{-4}	0.2041	0.0275	0.2659	3.62×10^{-4}	0.3043	0.0810	0.2719
TV-SDL	1.90×10^{-4}	0.1903	0.0403	0.1321	9.54×10^{-4}	0.2814	0.1960	0.1381
Non-TV	5.16×10^{-1}	0.4800	0.4800	0.5289	2.24×10^{-1}	0.7700	0.1500	0.5962

	$\mathbf{x}_i \overset{\text{iid}}{\sim} N(\mathbf{0}, I_p)$, $e \sim$ LRD				$\mathbf{x}_i \overset{\text{iid}}{\sim} N(\mathbf{0}, T)$, $e \sim$ LRD			
	with $\varrho = 0.75$				with $\varrho = 0.75$			
Method	FP(%)	FN(%)	FWER	RMSE	FPR	FNR	FWER	RMSE
TV-Lasso	7.25×10^{-2}	0.0652	1	0.0499	1.77×10^{-1}	0.0805	1	0.0556
FP-Lasso	1.60×10^{-4}	0.2496	0.0450	0.1091	1.37×10^{-4}	0.5138	0.0229	0.1381
Proposed	1.60×10^{-4}	0.1783	0.0433	0.1806	1.37×10^{-4}	0.3376	0.0243	0.1953
TV-LDPE	1.08×10^{-4}	0.2067	0.0238	0.2653	3.68×10^{-4}	0.3648	0.0859	0.2691
TV-SDL	2.10×10^{-4}	0.1924	0.0501	0.1386	9.29×10^{-4}	0.3014	0.0186	0.1448
Non-TV	5.19×10^{-1}	0.4800	0.4800	0.5304	5.49×10^{-1}	0.4500	0.4100	0.5280

Table 3. Runtime per 10 simulations.

Method	Runtime (in minutes)
TV-Lasso	0.5
FP-Lasso	9.5
Proposed	13
TV-LDPE	$> 1{,}000$
TV-SDL	$> 1{,}000$
Non-TV	< 0.5

considerable temporal variation on small time scales, Chang and Glover (2010); Hutchison et al. (2013). In view of the high spatial resolution of fMRI data, brain networks of subjects at rest are believed to be structurally homogeneous with subtle fluctuations in some, but a small number of, connectivity edges, Kiviniemi et al. (2005); Hutchison et al. (2010). A popular approach to learn brain connectivity is the neighborhood selection procedure, Meinshausen and Bühlmann (2006). Therefore, high-dimensional TVCM with a small number of nonzero components is a natural approach to study the time-evolving sparse brain connectivity networks, in which the time-varying coefficients reflect the dynamic features of the corresponding edges in the networks. The neighborhood selection approach is an approximation to the full multivariate distributions while ignoring the correlation among the node-wise responses and this may cause certain power

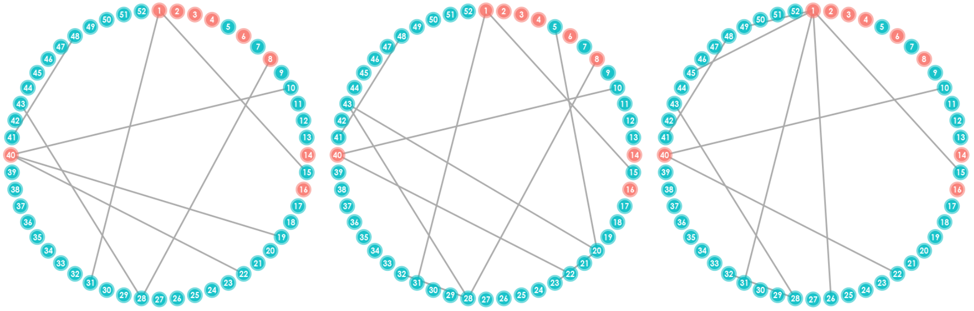


Figure 2. Connectivity networks in control subject around $t = 0.25$ based on the proposed method.

loss in finite samples. Meinshausen and Bühlmann (2006) showed that in terms of variable selection, these two approaches are asymptotically equivalent.

Our data example uses fMRI data collected from a study of patients with Parkinson's disease (PD) and their respective normal controls. PD is typically characterized by deviations in functional connectivity between various regions of the brain. Additionally, the resting state functional connectivity has been shown as a candidate biomarker for PD progression and treatment, where more advanced stages or manifestations of PD are associated with greater deviations from normal connectivity. Each resting state data matrix in our example contains 240 time points and 52 brain regions of interest (ROI). The time points are evenly sampled and the time indices are normalized to $[0, 1]$.

The brain connectivity network was constructed using the neighborhood selection procedure. In essence, it is a sequence of time-varying linear regressions enumerating each ROI as the response variable and sparsely regressing on all the other ROIs. Figure 2 and 3 show the estimated graphs of a normal subject and a PD subject at three sequential time points around $t = 0.25$ based on the proposed method. Red nodes are ROIs known to be associated with motor control and blue nodes are ROIs either known to be unrelated to motor control or whose functions in humans are not well understood. Different patterns of connectivity in the networks can be found by comparing normal and PD subjects. From the graphs, there are slow changes in the networks over time: most edges are preserved on a small time scale, but there are a few edges evolving over time. For instance, in a PD subject, ROI 1 and ROI 40 are unconnected in the first network but they are connected in the second network and remain connected in the third network.

We also plot the connectivity graphs generated by the competing methods

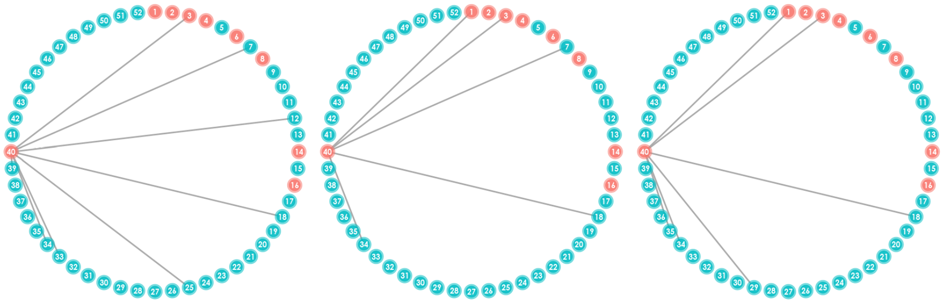


Figure 3. Connectivity networks in Parkinson's Disease subject around $t = 0.25$ based on the proposed method.

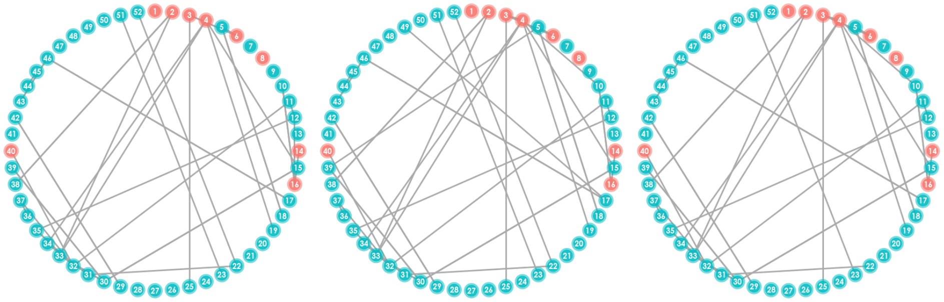


Figure 4. A connectivity network based on TV-LDPE.

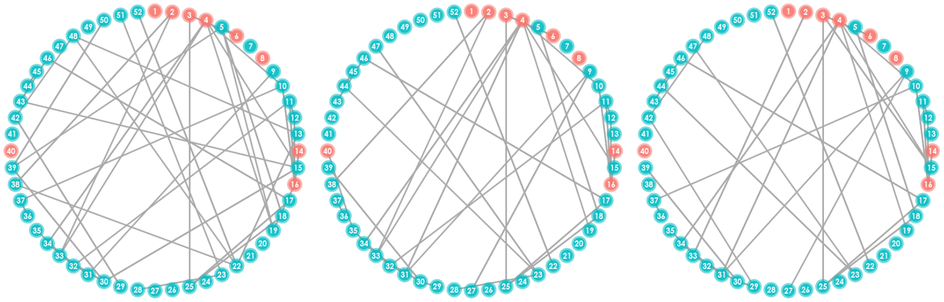


Figure 5. A connectivity network based on TV-SDL.

TV-LDPE and TV-SDL; see Figure 4 and 5. These graphs are denser than with the proposed method and are harder to interpret. Besides, as commented in the simulation studies, TV-LDPE and TV-SDL are more computationally expensive. The method of Bühlmann (2013) cannot be used to capture the dynamic features of brain connectivity networks.

6. Discussions

This paper presents a pointwise inference algorithm for high-dimensional

TVCM that can asymptotically control the FWER. Based on the current work, an interesting improvement would be to study simultaneous inference. Construction of the simultaneous confidence band (SCB) for the time-varying coefficients is useful for testing their parametric forms in high dimensions. This is a more challenging topic, which requires substantial additional work and probability tools that are beyond the scope of this paper.

Our brain connectivity application is a subject-by-subject analysis. To perform the group analysis on the population level, a hierarchical linear model is more appropriate. When p is fixed, the linear mixed-effects model is widely used in performing the multi-level group analysis in fMRI, Beckman, Jenkinson and Smith (2003); Lindquist (2008); Skup (2010). The reason is that the generalized least squares (GLS) estimator of a two-level model is inferentially equivalent to the GLS estimator of the corresponding single-level model, provided that the second-level covariance is the sum of the group covariance and the covariance of the first-level estimate, Beckman, Jenkinson and Smith (2003). Extension of the testing problem on the population parameters based on the ridge and Lasso estimates when $p \to \infty$ is another interesting future research topic.

7. Proof

Proof of Theorem 1. Observe that $\mathcal{X}_t\boldsymbol{\beta}(t) = \mathcal{X}_t\boldsymbol{\theta}(t)$ since $\boldsymbol{\theta}(t) = P_{\mathcal{R}_t}\boldsymbol{\beta}(t)$. Using the closed-form formulae for the tv-ridge estimator (2.6) and by (S0.4), we have

$$\begin{aligned}\text{bias}(\tilde{\boldsymbol{\theta}}(t)) &= \mathbb{E}(\tilde{\boldsymbol{\theta}}(t)) - \boldsymbol{\theta}(t)\\ &= (\mathcal{X}_t^\top\mathcal{X}_t + \lambda_2 I_p)^{-1}\mathcal{X}_t^\top[\mathcal{X}_t\boldsymbol{\theta}(t) + M_t\mathcal{X}_t\boldsymbol{\beta}'(t) + \mathcal{X}_t\boldsymbol{\xi}] - \boldsymbol{\theta}(t), \qquad (7.1)\end{aligned}$$

where $|\boldsymbol{\xi}|_\infty \le C_0 b_n^2/2$ and $|\boldsymbol{\xi}|_0 \le s$ almost surely, $t \in \varpi$. First, we bound the shrinkage bias of the tv-ridge estimator. By the argument in Section 3 of Shao and Deng (2012), we can show that

$$(\mathcal{X}_t^\top\mathcal{X}_t + \lambda_2 I_p)^{-1}\mathcal{X}_t^\top\mathcal{X}_t\boldsymbol{\theta}(t) - \boldsymbol{\theta}(t) = -Q(\lambda_2^{-1}D^2 + I_r)^{-1}Q^\top\boldsymbol{\theta}(t).$$

It follows from Lemma S0.2 that

$$\begin{aligned}|Q(\lambda_2^{-1}D^2 + I_r)^{-1}Q^\top\boldsymbol{\theta}(t)|_2 &\le \frac{|\boldsymbol{\theta}(t)|_2}{\rho_{\min}(\lambda_2^{-1}D^2 + I_r)} \qquad (7.2)\\ &= \left(\frac{\lambda_2}{\lambda_2 + \min_{j\le r} d_j^2}\right)|\boldsymbol{\theta}(t)|_2 \le \frac{\lambda_2|\boldsymbol{\theta}(t)|_2}{\rho_{\min\neq 0}(\hat{\Sigma}_t)} \le \frac{\lambda_2|\boldsymbol{\theta}(t)|_2}{|N_t|\underline{w}_t\varepsilon_0^2},\end{aligned}$$

where $d_j^2 = \rho_j(\hat{\Sigma}_t), j = 1, \cdots, r$. Next, we deal with the non-stationary bias of the tv-ridge estimator (7.1) by a similar argument for (7.2). Indeed, let $Q_\perp$ be the orthogonal complement of Q such that $Q_\perp^\top Q_\perp = I_{p-r}$ and $Q_\perp^\top Q = \mathbf{0}_{(p-r)\times r}$. Denote $\Gamma = [Q; Q_\perp]$. Then, $\Gamma\Gamma^\top = \Gamma^\top\Gamma = I_p$. By the SVD of $\mathcal{X}_t$, (2.2), we have

$$
\begin{aligned}
&(\mathcal{X}_t^\top \mathcal{X}_t + \lambda_2 I_p)^{-1}\mathcal{X}_t^\top M_t \mathcal{X}_t \boldsymbol{\beta}'(t) = \Gamma\left(\Gamma^\top (QD^2Q^\top + \lambda_2 I_p)\Gamma\right)^{-1}\Gamma^\top \mathcal{X}_t^\top M_t \mathcal{X}_t \boldsymbol{\beta}'(t)\\
&= [Q; Q^\perp]\left(\begin{bmatrix} Q^\top \\ Q_\perp^\top \end{bmatrix}(QD^2Q^\top + \lambda_2 I_p)[Q; Q_\perp]\right)^{-1}\begin{bmatrix} Q^\top \\ Q_\perp^\top \end{bmatrix} QDP^\top M_t \mathcal{X}_t \boldsymbol{\beta}'(t)\\
&= [Q; Q^\perp]\begin{pmatrix} (D^2 + \lambda_2 I_r)^{-1} & \mathbf{0} \\ \mathbf{0} & \lambda_2^{-1} I_{p-r} \end{pmatrix}\begin{bmatrix} DP^\top M_t \mathcal{X}_t \boldsymbol{\beta}'(t) \\ \mathbf{0} \end{bmatrix}\\
&= Q(D + \lambda_2 D^{-1})^{-1} P^\top M_t \mathcal{X}_t \boldsymbol{\beta}'(t).
\end{aligned}
$$

Hence, by Lemma S0.2 we have

$$
|(\mathcal{X}_t^\top \mathcal{X}_t + \lambda_2 I_p)^{-1}\mathcal{X}_t^\top M_t \mathcal{X}_t \boldsymbol{\beta}'(t)|_2 \le \frac{b_n |\mathcal{X}_t \boldsymbol{\beta}'(t)|_2}{\rho_{\min}(D + \lambda_2 D^{-1})} \le \frac{C_0 b_n (s|N_t|\overline{w}_t)^{1/2}\varepsilon_0^{-1}}{\min_{j\le r}(d_j + \lambda_2/d_j)},
$$

where $\overline{w}_t = \sup_{i\in N_t} w(i,t)$. Since $\lambda_2 = o(1)$ and $d_j \ge (|N_t|\underline{w}_t)^{1/2}\varepsilon_0$, the denominator of last expression is lower bounded by $[(|N_t|\underline{w}_t)^{1/2}\varepsilon_0 + \lambda_2/((|N_t|\underline{w}_t)^{1/2}\varepsilon_0)]$ for large enough n. Therefore, we have

$$
|(\mathcal{X}_t^\top \mathcal{X}_t + \lambda_2 I_p)^{-1}\mathcal{X}_t^\top M_t \mathcal{X}_t \boldsymbol{\beta}'(t)|_2 \le \frac{C_0 b_n (s|N_t|\overline{w}_t)^{1/2}}{(|N_t|\underline{w}_t)^{1/2}\varepsilon_0^2} \le \frac{C_0 b_n s^{1/2}}{C\varepsilon_0^2}. \tag{7.3}
$$

Similarly, an upper bound for the remainder term of (7.1) can be established. We have

$$
|(\mathcal{X}_t^\top \mathcal{X}_t + \lambda_2 I_p)^{-1}\mathcal{X}_t^\top M_t \mathcal{X}_t \boldsymbol{\xi}|_2 \le \frac{C_0 b_n^2 s^{1/2}}{2C\varepsilon_0^2}, \quad \text{for almost surely } t \in \varpi. \tag{7.4}
$$

In addition, $\tilde{\boldsymbol{\theta}}(t) - \mathbb{E}[\tilde{\boldsymbol{\theta}}(t)] = (\mathcal{X}_t^\top \mathcal{X}_t + \lambda_2 I_p)^{-1}\mathcal{X}_t^\top \mathcal{E}_t$ is the stochastic part of the tv-ridge estimator. Since the $e_i \sim N(0, \sigma^2 I_n)$ are iid, $\mathcal{E}_t \sim N(\mathbf{0}, \sigma^2 W_t)$. Hence, $\tilde{\boldsymbol{\theta}}(t) - \mathbb{E}[\tilde{\boldsymbol{\theta}}(t)] \sim N(\mathbf{0}, \Omega(\lambda_2))$, where $\Omega(\lambda)$ is defined in (2.4), and thus

$$
\operatorname{Var}(\tilde{\theta}_j(t)) = \Omega_{jj}(\lambda_2) \ge \Omega_{\min}(\lambda_2). \tag{7.5}
$$

Now, we consider the initial tv-lasso estimator. By Lemma S0.3,

$$
|\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)|_1 \le 4\phi_0^{-2}\lambda_1 s. \tag{7.6}
$$

Then, (3.6), (3.7), and (3.8) follow by assembling (7.2), (7.3), (7.4), and (7.6) into (2.7),

$$
\hat{\boldsymbol{\beta}}(t) = \boldsymbol{\beta}(t) + \operatorname{bias}(\tilde{\boldsymbol{\theta}}(t)) + \{\tilde{\boldsymbol{\theta}}(t) - \mathbb{E}[\tilde{\boldsymbol{\theta}}(t)]\} - \{(P_{\mathcal{R}_t} - I_p)[\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)]\}.
$$

The marginal representation (3.9) and (3.10) follow from similar arguments by noting that $B_j(t) = \sum_{k\ne j}(P_{\mathcal{R}_t})_{jk}\beta_k(t)$ under $H_{0,j,t}$.

Proof of Theorem 3. The proof of Theorem 3 is similar to that of Theorem 1 so we only highlight the difference involving the error process. First, $\mathrm{Cov}(\mathcal{E}_t) = W_t^{1/2}\Sigma_{e,t}W_t^{1/2}$. Second, instead of using (S0.5) in proving Lemma S0.3, we use Lemma S0.4 to get for all $\lambda > 0$

$$\mathbb{P}\left(\max_{j\le p}\left|\sum_{i\in N_t} w(i,t)X_{ij}e_i\right| \ge \lambda\right) \le 2p\exp\left(-\frac{\lambda^2}{2L_{t,2}^2|\mathbf{a}|_1^2\sigma^2}\right) \quad \text{if } \varrho > 1,$$
$$\mathbb{P}\left(\max_{j\le p}\left|\sum_{i\in N_t} w(i,t)X_{ij}e_i\right| \ge \lambda\right) \le 2p\exp\left(-\frac{C_\varrho\lambda^2}{L_{t,2}^2 n^{2(1-\varrho)}\sigma^2K^2}\right) \quad \text{if } \varrho \in (\frac{1}{2},1).$$

Proof of Theorem 4. The proof essentially follows the lines of that of Theorem 1, but with differences in requiring a larger penalty parameter λ_1 of the tv-Lasso. First, by the Nagaev inequality (Nagaev (1979)), we have for any $\varepsilon > 0$,

$$\mathbb{P}\left(\max_{j\le p}\left|\sum_{i\in N_t} w(i,t)X_{ij}e_i\right| \ge \sigma L_{t,2}\varepsilon\right) \le \left(1+\frac{2}{q}\right)^q \kappa_q\frac{p\mu_{n,q}}{(\sigma L_{t,2}\varepsilon)^q} + 2p\exp\left(-c_q\varepsilon^2\right),$$

where $c_q = 2e^{-q}(q+2)^{-2}$ and κ_q is the q-th absolute moment of e_1. Then, choosing

$$\varepsilon = C_q\max\left\{\frac{(p\mu_{n,q})^{1/q}}{\sigma L_{t,2}},\ (\log p)^{1/2}\right\} \quad \text{for large enough } C_q > 0,$$

we have $\max_{j\le p}|\sum_{i\in N_t} w(i,t)X_{ij}e_i| = O_{\mathbb{P}}(\lambda_0)$. Second, let $\Xi = (\mathcal{X}_t^\top\mathcal{X}_t + \lambda_2 I_p)^{-1}\mathcal{X}_t^\top W_t^{1/2}$ and $\mathcal{E}_t^\diamond = (e_i)_{i\in N_t}^\top$. Recall that $\tilde{\boldsymbol{\theta}}(t) - \mathbb{E}[\tilde{\boldsymbol{\theta}}(t)] = \Xi\mathcal{E}_t^\diamond$. By the Gaussian approximation (Shao, 1995, Thm. B), there exist iid Gaussian random variables $g_i \sim N(0,\sigma^2\xi_{ji}^2)$ defined on a possibly richer probability space such that for every $t > 0$

$$\mathbb{P}\left(\left|\tilde{\theta}_j(t) - \mathbb{E}[\tilde{\theta}_j(t)] - \sum_{i\in N_t} g_i\right| \ge t\right) \le (Cq)^q\frac{\sum_{i\in N_t}\mathbb{E}|\xi_{ji}e_i|^q}{t^q}.$$

Thus, it follows that $\tilde{\theta}_j(t) - \mathbb{E}[\tilde{\theta}_j(t)] = N(0,\Omega_{jj}(\lambda_2)) + O_{\mathbb{P}}(|\boldsymbol{\xi}_j|_q)$. As $\Omega_{jj}(\lambda_2) = \sigma^2|\boldsymbol{\xi}_j|_2^2$, the proof is complete for by assumption, $|\boldsymbol{\xi}_j|_q = o(|\boldsymbol{\xi}_j|_2)$.

Supplementary Materials

The supplementary material contains additional technical lemmas and discusses some implementation issues.

Acknowledgment

We would like to thank an anonymous referee, an Associate Editor, and the Co-Editor for their many helpful comments that led to improvements of this paper. This research was partially supported by NSF DMS-1404891 and UIUC

Research Board Award RB15004. We thank Aiping Liu (University of British Columbia) for providing the Parkinson's Disease data. The high-performance computing of this work was completed on the Illinois Campus Cluster Program at UIUC.

References

Beckman, C. F., Jenkinson, M. and Smith, S. M. (2003). General multilevel linear modeling for group analysis in FMRI. *NeuroImage* **20**, 1052–1063.

Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics* **37**, 1705–1732.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19**(4), 1212–1242.

Cai, Z. (2007). Trending time-varying coefficient time series models with serially correlated errors. *Journal of Econometrics* **136**, 163–188.

Chang, C. and Glover, G. H. (2010). Time-frequency dynamics of resting-state brain connectivity measured with fMRI. *NeuroImage* **50**, 81–98.

Chen, X., Xu, M. and Wu, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *Annals of Statistics* **41**, 2994–3021.

Chen, X., Xu, M. and Wu, W. B. (2016). Regularized estimation of linear functionals of precision matrices for high-dimensional time series. *IEEE Transactions on Signal Processing* **64**, 6459–6470.

Cleveland, W. S., Grosse, E. and Shyu, W. M. (1991). Local regression models. *In Statistical Models in S (Chambers, J. M. and Hastie, T. J., eds) Wadsworth & Brooks, Pacific Grove.* pp. 309–376.

Fan, J. and Han, X. (2016). Estimation of the false discovery proportion with unknown dependence, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* .

Fan, J., Han, X. and Gu, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *Journal of American Statistical Association* **107**, 1019–1035.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *Annals of Statistics* **27**, 1491–1518.

Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809–822.

Hutchison, M., S., M., Jones, C., Gati, J. and Leugn, L. (2010). Functional networks in the anesthetized rat brain revealed by independent component analysis of resting-state fMRI. *J. Neurophysiol* **103**, 3398–3406.

Hutchison, M., Womelsdorf, T., Gati, J., Everling, S. and Menon, R. (2013). Resting-state networks show dynamic functional connectivity in awake humans and anesthetized macaques. *Human Brain Mapping* **34**, 2154–2177.

Javanmard, A. and Montanari, A. (2014). Confidence inter hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory* **60**, 6522–6554.

Kiviniemi, V., Haanpää, H., Kantola, J.-H., J., J., V., V., Alahuhta, S. and Tervonen, O. (2005).

Midazolam sedation increases fluctuation and synchrony of the resting brain BOLD signal. *Magn Reson Imaging* **23**, 531–537.

Lindquist, M. (2008). The statistical analysis of fmri data. *Statistical Science* **23**, 439–464.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics* **34**, 1049–1579.

Nagaev, S. (1979). Large deviations of sums of independent random variables. *Annals of Probability* **7**, 745–789.

Orbe, S., Ferreira, E. and Rodriguez-Poo, J. (2005). Nonparametric estimation of time varying parameters under shape restrictions. *Journal of Econometrics* **126**, 53–77.

Robinson, P. M. (1989). *Nonparametric estimation of time-varying parameters*, In Statistical Analysis and Forecasting of Economic Structural Change (P. Hackl, ed.) Berlin: Springer.

Shao, J. and Deng, X. (2012). Estimation in high-dimensional linear models with deterministic design matrices. *Annals of Statistics* **40**, 812–831.

Shao, Q.-M. (1995). Strong approximation theorems for independent random variables and their applications. *Journal of Multivariate Analysis* **52**, 107–130.

Skup, M. (2010). Longitudinal fmri analysis: a review of methods. *Statistics and Its Interface* **3**, 235–252.

Song, R., Yi, F. and Zou, H. (2014). On varying-coefficient independence screening for high-dimensional varying-coefficient models. *Statistica Sinica* **24**, 1735–1752.

van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics* **3**, 1360–1392.

Wei, F., Huang, J. and Li, H. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica* **21**, 1515–1540.

Xue, L. and Qu, A. (2012). Variable selection in high-dimensional varying coefficient models with global optimality. *Journal of Machine Learning Research* **13**, 1973–1998.

Zhang, C.-H. and Zhang, S. S. (2013). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**, 217–242.

Zhang, T. and Wu, W. B. (2012). Inference of time-varying regression models. *Annals of Statistics* **40**, 1376–1402.

Zhang, W., Lee, S.-Y. and Song, X. (2002). Local polynomial fitting in semivarying coefficient model. *Journal of Multivariate Analysis* **82**, 166–188.

Zhou, S., Lafferty, J. and Wasserman, L. (2010). Time-varying undirected graphs. *Machine Learning* **80**, 295–319.

Zhou, Z. and Wu, W. B. (2010). Simultaneous inference of linear models with time varying coefficients. *Journal of the Royal Statistical Society* **72**, 513–531.

Department of Statistics, University of Illinois at Urbana-Champaign, 725 S Wright Street, Champaign, IL USA 61820

E-mail: xhchen@illinois.edu

Department of Statistics,University of Illinois at Urbana-Champaign, 725 S Wright Street, Champaign, IL USA 61820

E-mail: he3@illinois.edu

(Received June 2015; accepted December 2016)